DaSe Lab

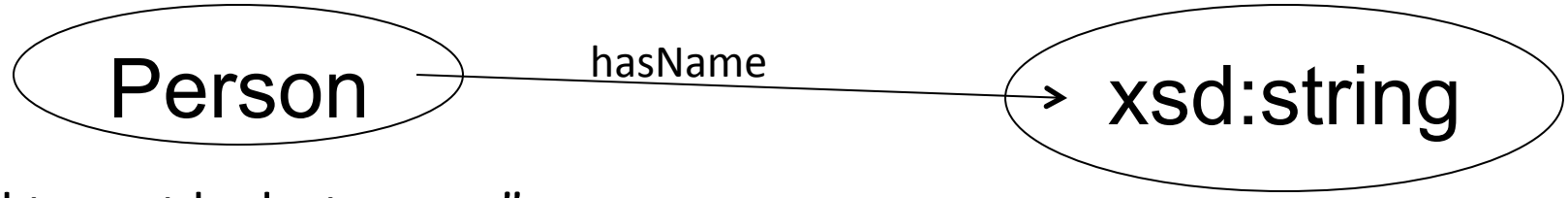
Ontology Design Patterns:
Piecing together an introduction

Michelle Cheatham and Pascal Hitzler

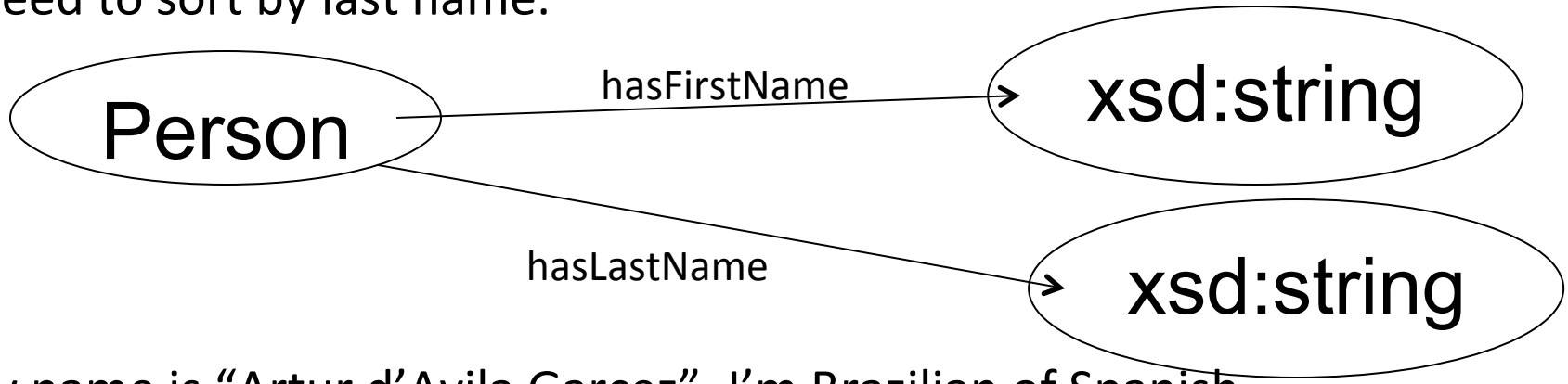
ESIP Semantic Web Telecon
September 2015

WRIGHT STATE
UNIVERSITY

Names



"I need to sort by last name."



"My name is "Artur d'Avila Garcez". I'm Brazilian of Spanish descendancy.

"My first name is Anna-Maria, but I live in the U.S. and the ID systems didn't accept a hyphen in my name.

"My name is Pan Ji. What do you mean by 'last name'?"

"My name actually changed recently ..."

Ontological Commitments

a.k.a.

modeling choices you may regret later

Ontological Commitments

Whenever you decide on how to make your metadata

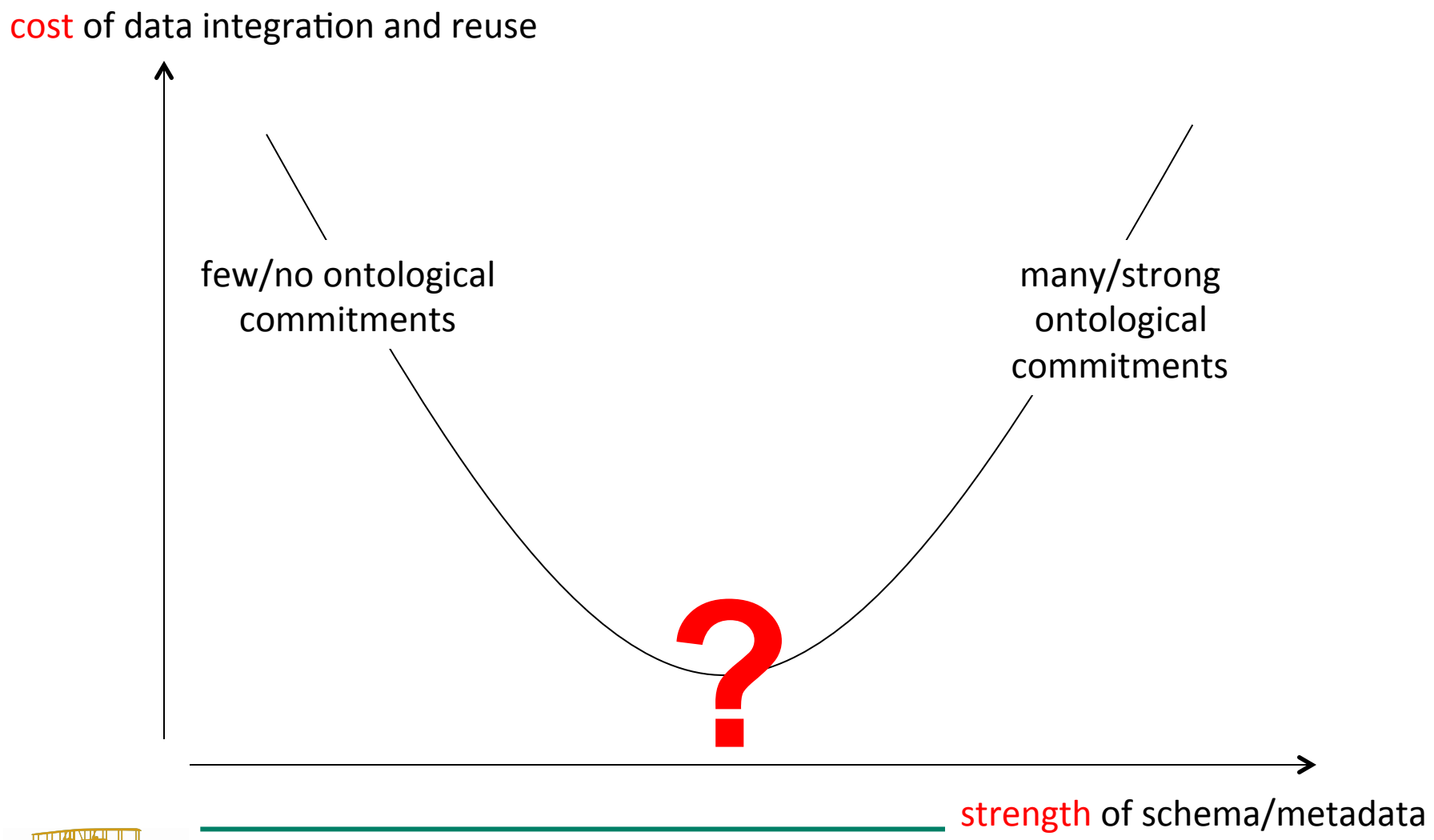
- keyword annotation
- controlled vocabularies
- light-weight taxonomy
- full-blown ontology

You always have to make specific modeling decisions.

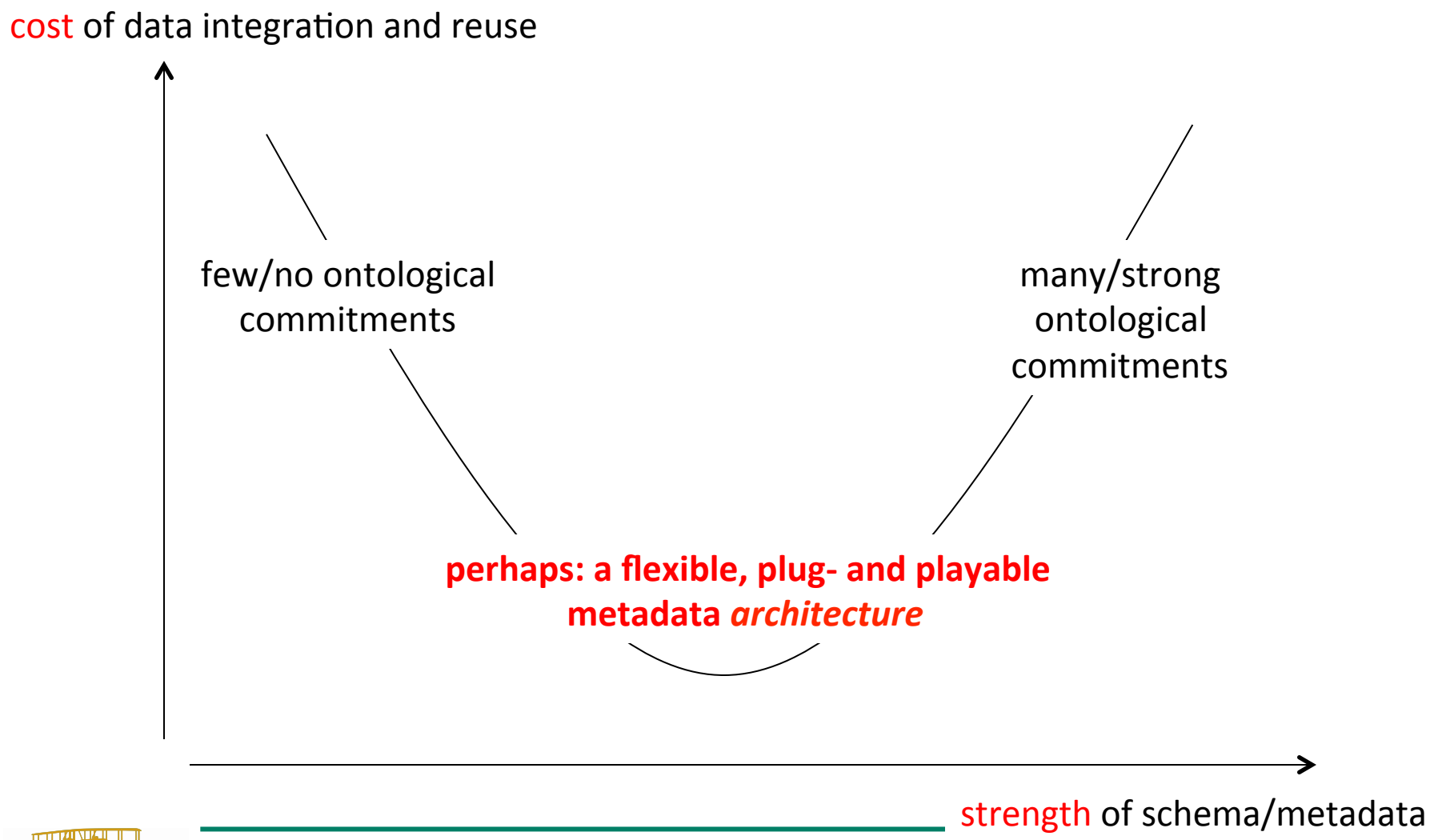
You can either make detailed specifications (ontological commitments) which will often hinder reuse for new purposes.

Or you can avoid the commitments, resulting in ambiguity which cannot really be resolved later, thus also hindering reuse.

Soft Spot Search



Soft Spot Search



Ontology Design Patterns

"An ontology design pattern is a reusable successful solution to a recurrent modeling problem."

So-called *content patterns* usually encode specific abstract notions, such as process, event, agent, etc.

Patterns provide modular, reusable, replaceable, pieces.

Patterns can be configured as a flexible, modular, "plug-and-play" metadata ecosystem in which patterns can be exchanged as needed.

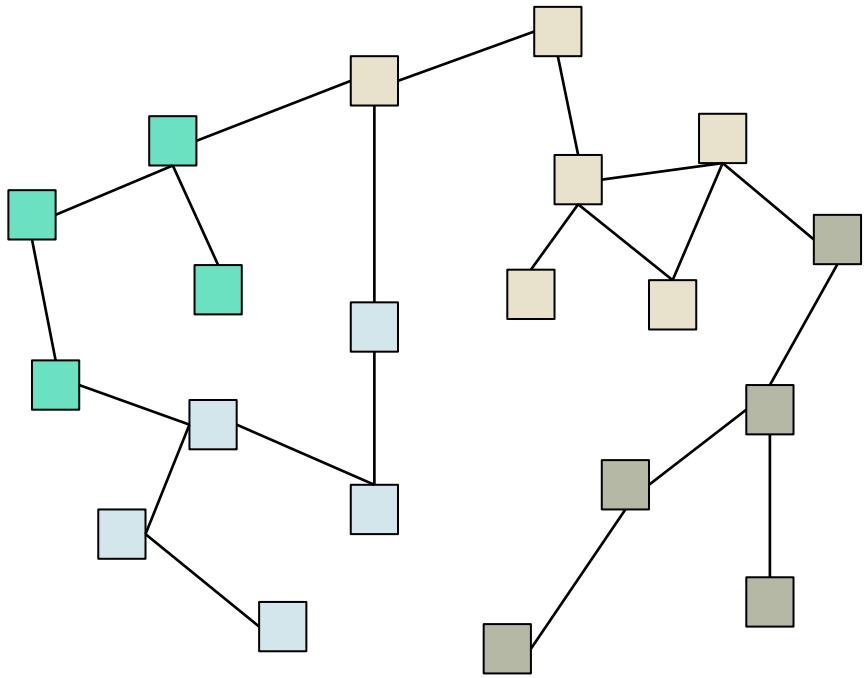
ODPs versus ontologies

The complexity and scope of knowledge that man has gained about various aspects of the physical world has led to scientists to specialize further and further, to the point where some researchers now spend their entire careers studying the behavior of one particular protein or a single species of plankton. This level of focus is unavoidable to some degree – no one person can be an expert in everything. However, it is sometimes easy to forget that the world around us is an interconnected system whose behavior spans the artificial boundaries between traditional scientific disciplines, and often the greatest leaps forward in our understanding come at the intersection of different areas of study. These types of breakthroughs require the integration of data from many different scientific domains, and this integration must be done in such a way that the detail, uncertainty, and above all the *context* of the data are preserved.

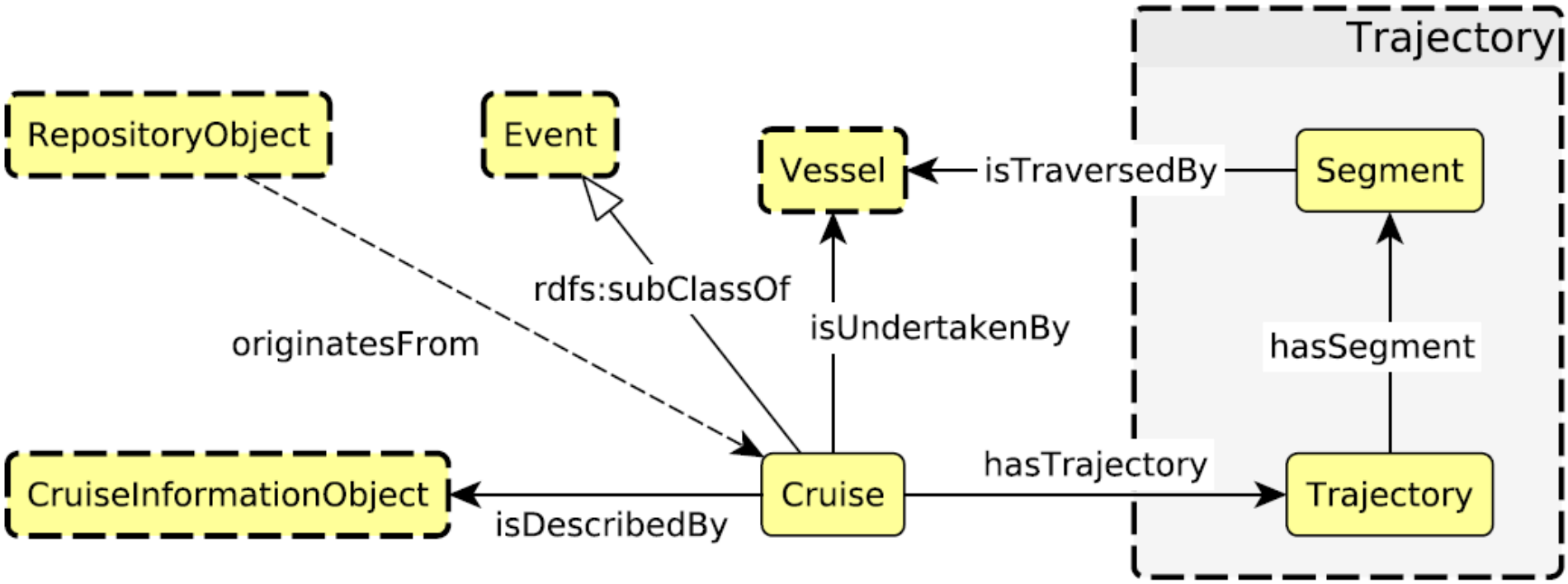
The first challenge of scientific data integration is accessibility. Much of the data underpinning past and present scientific publications is not readily accessible – it exists only in isolated databases, as files on a grad student's computer, or in tables within PDF documents. The consequence of this is that it is often difficult to replicate published experimental results and to do new analyses on existing data. There is a monetary cost as well: if data is not stored and shared in an accessible manner then it must be collected independently by multiple researchers, using limited scientific funding that could better be spent elsewhere. Fortunately, many funding agencies have taken action on this issue and now require that data collected via funded programs be stored in official data repositories. This is a promising development, but many current data repositories do not make data integration easy. Traditional scientific data repositories are generally either relational databases or file servers containing spreadsheet, CSV, or irregularly structured text files. There can be various obstacles to retrieving this data, particularly due to a lack of consistency. For instance, some repositories might be accessible via websites or structured query mechanisms while others require a login and use of secure file transfer or copy protocols. Financial and legal concerns also inhibit data integration. Some data might be stored in proprietary database or file formats that require expensive software licenses to read, and licenses indicating what users are allowed to do with the data can be missing or restrictive, resulting in legal uncertainty.

One approach that has been proposed to address these accessibility issues is publishing data as "linked data." In a linked dataset, every entity is given a URI that can be accessed via HTTP, similar to a standard webpage. When this URI is dereferenced, there is structured data providing more information about the entity. This data is generally expressed using RDF, and it may include URIs to other, related, entities. Using these standards, it possible to make data available in a way that is both accessible and understandable.

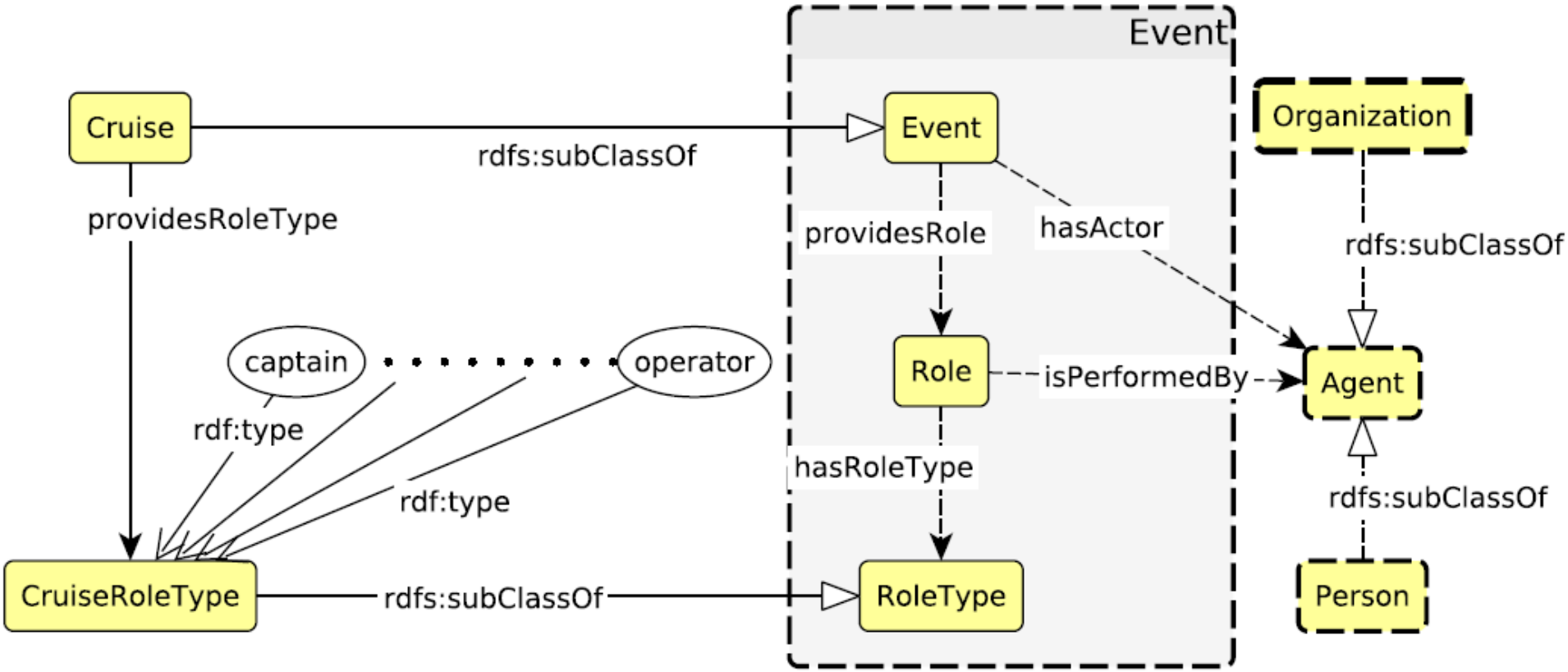
An example to clarify this description seems warranted. Assume that Dr. Jane Doe, a scientist at State University, wants to publish a linked dataset containing information about the papers she has written, one of which is called "An Exploration of the Feasibility of Tenure." One way for Dr. Doe to do this is to acquire ownership of a domain name and assign URIs in that namespace to each of the entities in her dataset. For instance, if the domain name is profdoe.edu then she might use the URI profdoe.edu/JaneDoe to represent herself and profdoe.edu/TenureFeasibilityExploration to represent the paper. Dr. Doe could then create files containing RDF statements about these entities and deploy them on a webserver. An RDF statement is a subject-predicate-object triple. For example, the following triple states that the paper's title is "An Exploration of the Feasibility of Tenure":



Oceanographic Cruise



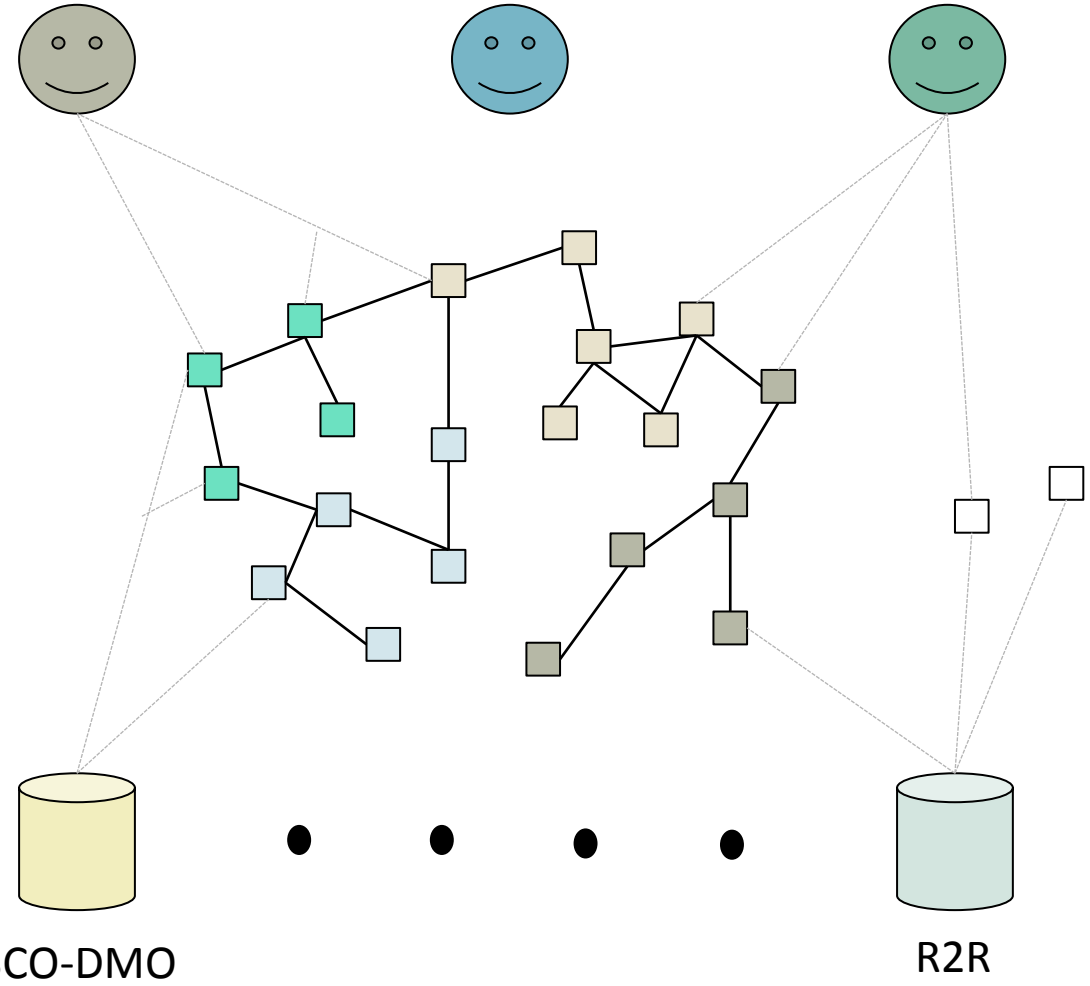
Cruise as Event



GeoLink Patterns

- Person
- Place
- Organization
- Agent Role
- Cruise
- Vessel
- Funding Award
- Program
- Information Object
- Physical Sample
- Measurement
- Dataset
- Publication/Document

GeoLink setup



Benefits of ODPs

- Data integration, particularly when the datasets
  - are very large
  - have a spatiotemporal aspect
  - involve different sensor modalities
  - use very different measurement scales
  - have only a small area of overlap

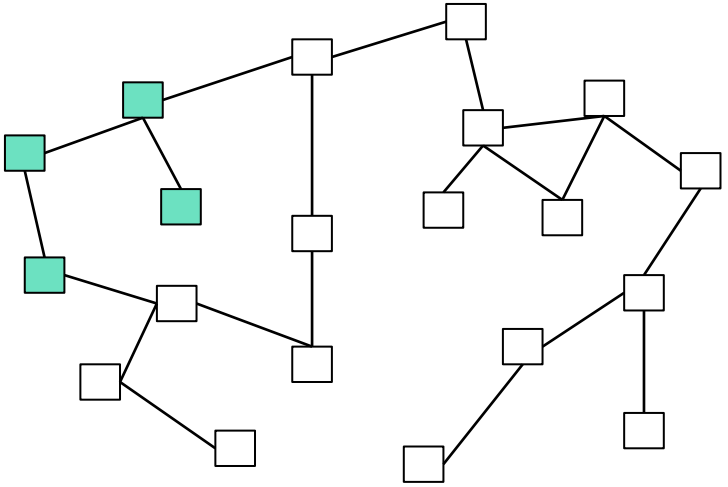
- ODPs provide a structured and application-neutral representation of the key concepts within a domain.

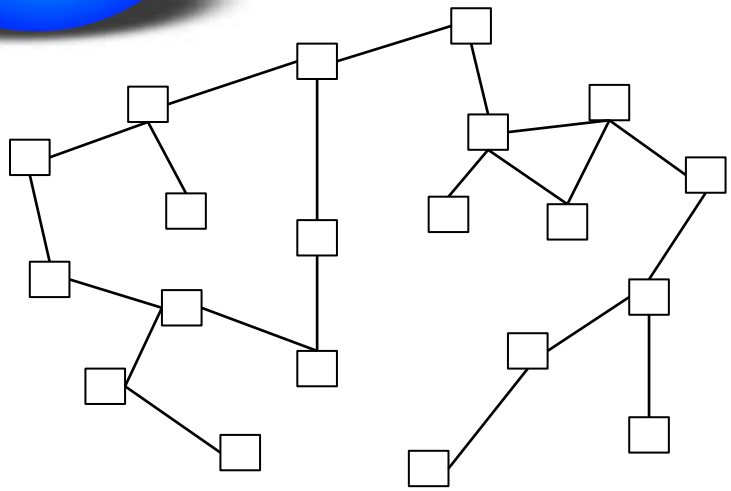
- These are frequently the small areas of semantic overlap that exist between datasets from different subfields of the same high-level domain.

- Diverse data can be integrated and queried across without the need to shoehorn all the data from both into the same overarching structure

Benefits of ODPs

?





!

Thank you!

Questions?